DOCUMENT RESUME

ED 307 321 TM 013 463

AUTHOR	Auchter, Joan Chikos; Patience, Wayne
TITLE	Decentralized Large Scale Essay Scoring: Methods for Establishing and Evaluating Score Scale Stability and Reading Reliability.
PUB DATE	Mar 89
NOTE	41p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, March 28-30, 1989).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Decentralization; Equivalency Tests; *Essay Tests; Evaluators; High School Equivalency Programs; Holistic Evaluation; Interrater Reliability; *Scoring; Secondary Education; Testing Problems; *Testing Programs; Writing (Composition); *Writing Evaluation
IDENTIFIERS	Direct Assessment; *General Educational Development Tests; *Large Scale Programs; Score Variation

ABSTRACT

The methods used by the General Educational Development Testing Service (GEDTS) to establish and maintain score stability and reading reliability on its direct assessment of writing are described. Using the 1988 site certification and monitoring results of several scoring sites, the focus is on describing how the score scale was established and how reader training and certification, scoring site certification, and scoring monitoring procedures were developed and implemented. The Tests of General Educational Development (GED) are used extensively to award high school equivalency diplomas. A writing essay was added in the 1988 administration. The nature of the tests required that essays be scored in decentralized scoring sessions. Holistic scoring with two readers was chosen, with a third reading if scores differed too much. The GED direct assessment is norm-referenced, calling for a descriptive, rather than prescriptive, scoring guide. Chief readers and readers receive training in scoring, and a site is only certified when readers have been trained. Systematic monitoring reduces or eliminates scoring drift. Approximately 150 persons attended the first Chief Reader training. At the end of the first year of site certification, 52 operational scoring sites were established. Monitoring results of 15 sites are presented. Ensuring fair and stable scoring from site to site requires careful planning and systematic procedures. The process so far illustrates the importance of training readers and monitoring the sites. Seven tables present monitoring results. Nine appendices provide additional information about the development process, including scoring, test instructions, and reader requirements. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED307321

Decentralized Large Scale Essay Scoring: Methods for Establishing and Evaluating Score Scale Stability and Reading Reliability

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Joan Chikos Auchter and Wayne Patience

GED Testing Service
American Council on Education

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JOAN AUCHTER
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Abstract

A central aim in any assessment program is to ensure fair and stable scoring from administration to administration. When administrations are decentralized, not only in location, but also in frequency, and in logistical configuration, it is imperative to construct training, certifying, and monitoring systems which provide continuity between the original application of scoring standards and the application of those standards from testing site to testing site and from test administration to test administration. The 1988 Tests of General Educational Development (GED Tests) include a direct measure of writing as part of the Writing Skills Test. Much of the GED Testing Service's research on the feasibility of including an essay focused on defining and maintaining a holistic score scale and on identifying methods of ensuring score scale stability and reading reliability across scoring sessions and decentralized sites. Because the Writing Skills Test scores, along with the scores on the other four subject area tests in the battery, determine whether an examinee receives a high school equivalency credential, the reliability of the essay and multiple-choice composite score on the Writing Skills Test is of major importance. This paper will describe the methods of establishing and monitoring score scale stability and reading reliability and will report how well these methods have succeeded in a decentralized essay scoring program.

Paper presented at the annual meeting of The National Council on Measurement in Education, San Francisco, California, March 1989.

TM013463

BEST COPY AVAILABLE

Decentralized Large Scale Essay Scoring: Methods for Establishing and Evaluating Score Scale Stability and Reading Reliability

Joan Chikos Auchter and Wayne Patience
GED Testing Service

The primary purpose of this presentation is to describe and illustrate the methods employed by the General Educational Development Testing Service (GEDTS) to establish and maintain score stability and reading reliability. More specifically, the focus is on describing how the score scale was established and how reader training and certification, scoring site certification, and scoring monitoring procedures were developed and implemented. Using the 1988 site certification and monitoring results of several scoring sites, the paper describes how effective these methods were during the first full year of the decentralized scoring of the GED direct assessment of writing. Before turning to this focus, some background is required to understand the context in which choices were made in developing the 1988 GED Writing Skills Test.

Purpose of the Tests of General Educational Development (GED Tests)

GED Tests are used extensively by states, Canadian provinces, and U. S. and Canadian territories to award a high school equivalency diploma to those who did not complete a high school program. Nearly three-quarters of a million examinees are tested annually, and almost half a million people earn a GED credential each year. Approximately one in seven high school completion credentials issued in the United States is a GED credential.

The credential issued on the basis of the GED Test results is commonly referred to as a high school equivalency diploma. This notion of equivalency derives from: 1) defining the content of the GED Tests to reflect the commonly expected outcomes of completing a traditional high school program of study, and 2) defining passing scores relative to the actual demonstrated performance of contemporary graduating seniors.

The authority to award a high school equivalency diploma rests with the participating jurisdictions. In part, this explains the extremely decentralized nature of the GED Program. GED Tests are administered in over 3,000 Official GED Testing Centers across North America, and are given nearly every day of the year.

Adding an Essay to the GED Writing Skills Test

Periodically, the GED Tests are extensively reviewed to evaluate how well they reflect current high school curriculum and expectations. Such a review was begun in 1983 by surveying secondary school curriculum specialists, adult education teachers and program directors, and GED Administrators and Examiners (Spille, 1984). National secondary school associations were also polled, and, in addition to being asked to complete the surveys, they were asked to nominate individuals specializing in the major secondary school subject areas (English, Social Studies, Science, and Mathematics) who would be well qualified to serve

on a GED Tests Specifications Committee. Committee members were selected from these nominations. In February 1984, after having reviewed the current GED Tests and many reports regarding secondary school curriculum, the Committee met. In June of 1984, the Committee released its GED Tests Specifications Report (GEDTS Tests Specification Committee, 1984). The addition of an essay to the GED Writing Skills Test was among the primary recommendations of the GED Test Specifications Committee.

After reviewing the research on large scale writing assessments and conducting research on: 1) the relationship of the current multiple-choice Writing Skills Test and direct measures of writing skills (GED Testing Service, 1985a), and 2) the reliability and validity issues of using a direct measure of writing as part of the GED Tests (GEDTS, 1985b), the GED Testing Service staff recommended the incorporation of an essay to the Commission on Educational Credit and Credentials, the independent policy making body for the GED Testing Service. With attention to the views of participating jurisdictions, the Commission ultimately approved the addition of an essay at its fall 1985 meeting.

Reliability Issues Associated with Adding the Essay

Among the problems which contribute to the relatively low score reliability of direct measures of writing are: controlling score variability due to topic variability, and maximizing reader consistency in applying a score scale. Research has suggested that to attain comparable score reliability for a direct measure of writing to that of a multiple-choice test, an essay test would require at least six essays, each scored by at least two readers (Breland, 1986). Early on, it was evident that, given the costs associated with scoring essays, only one essay could be given as part of the GED Tests. Consequently, it was clear that the essay would have to be incorporated with a multiple-choice section in order to obtain an acceptably valid and reliable measure of writing skills. Given the decentralized nature of the GED Testing Program (many testing sites and testing dates) and the traditional values of local autonomy and rapid return of test results to examinees, it was also clear that it would not be feasible to require all 3,000 Official GED Testing Centers to send essays to a GED Scoring Service for scoring. Therefore, GEDTS offered GED Administrators four possible scoring configurations: 1) a central scoring site in the jurisdiction, 2) multiple scoring sites in the jurisdiction, 3) a commercial scoring site, or 4) the GED Essay Scoring Service.

Developing and Maintaining Score Scale Stability

Because the nature of the GED Testing Program requires that essays be scored in decentralized scoring sessions, a major concern is the consistency with which essay readers apply the scoring guide standards. In research comparing reading reliability and score scale stability across scoring sites (GEDTS, 1985b), it became evident that, while respectable reading reliability and score scale stability could be attained, rigorous procedures for controlling and monitoring essay scoring at the various scoring sites would have to be applied. Among the

procedures critical to obtaining high reading reliability and maintaining scale stability are: 1) defining the purpose of the direct writing assessment and developing an integrated model, 2) selecting the scoring rubric, 3) developing strong scoring standards and a scoring guide, 4) evaluating and selecting uniform topics, 5) developing standards, procedures, and materials for the training and certification of Chief Readers and readers (see Appendix A for a discussion on reader training), 6) designing, developing, and implementing site certification standards, procedures, and materials that evaluate whether a potential site is performing well enough to be certified as an official site to score operational essays, and 7) designing, developing, and implementing scoring site monitoring standards, procedures, and materials that evaluate whether the site's ongoing performance is sufficient to warrant continued certification to score operational essays.

Defining the Purpose

There are as many programs for scoring direct samples of writing as there are purposes for the programs (i.e., programs to diagnose writing, to promote excellence in writing, and to serve as a criterion reference.) Different purposes will often dictate different choices at various stages of developing the essay scoring model. The first and most crucial step in the development of any essay scoring model or program is the identification and articulation of the purpose for the direct writing assessment--form will then follow function. The program must be integrated from the onset.

Because the purpose of the GED Tests is to compare the educational achievement of the GED candidates with the demonstrated performance of contemporary graduating seniors, the purpose of the direct assessment was already established. The evaluation of the essay component of the Writing Skills Test would be norm-referenced--designed to compare the performance of GED candidates with the performance of a national stratified random sample of graduating high school seniors.

Selecting the Scoring Rubric

Of the scoring rubrics investigated, holistic scoring was chosen as the method of evaluating GED examinee essays. An explanation of the rationale for choosing holistic scoring is given as Appendix B (excerpted from GEDTS, 1985a). The scale used is a six point scale ranging from one (low) to six (high). The two readers scores are then added, resulting in a range of scores from two to twelve. If the two readers' scores differ by more than one point, the essay is scored a third time by the Chief Reader. The total score for essays scored by three readers is twice the average of the three scores.

After making the programmatic decisions of purpose and scoring rubric, the next and most essential task in developing and implementing a program that requires that 750,000 essays be scored annually in approximately 55

decentralized scoring sites and hundreds of scoring sessions, is the establishment of strong scoring standards. GEDTS has employed the following strategies to insure valid and reliable standards.

Establishing the GEDTS Writing Committee

The establishment of a permanent GEDTS Writing Committee was one of the most critical phases of the process. It was imperative to establish a group who would not only provide "expert judgement" in a variety of aspects for the direct writing assessment, but also maintain a historical perspective on the development and implementation of the essay component; providing continuity from administration to administration, from topic to topic, from scoring site to scoring site, and from testing session to testing session. This consistency in applying the scoring criteria to essays across administrations, topics, scoring sites, and scoring sessions is essential to ensuring that the papers are scored accurately over time.

The Writing Committee contributed to the establishment of score stability by assisting with the decision on the type of essay to be required, the choice of the scoring rubric, the development of the GED Essay Scoring Guide which serves as the criteria, and the selection of reader training paper sets.

The GED Writing Committee, composed of university, adult education, and high school instructors, was structured to represent all GED populations. Because the Committee members are directly involved in writing instruction on a day-to-day basis, they are familiar with instructional expectations for candidates at particular performance levels.

Ongoing roles of the Committee to maintain score scale stability include: 1) reviewing and evaluating potential topics, 2) participating in reading sessions designed to identify essays to be used in training and certification sets for Chief Readers and readers, 3) participating in reading sessions designed to identify essays to be used in site certification and monitoring sets, and 4) participating in reading sessions designed to identify an anchor set and a minimum of 12 sets of essays for each operational topic to be used to recalibrate readers to the GED Official Scoring Guide at the beginning of every official essay scoring session.

Development of the Official GEDTS Scoring Guide

To ensure clarity and a standard for evaluating the consistency of essay scoring across topics, as well as within and across scoring sites and sessions, a generic scoring guide was developed. Because the GED direct assessment is norm-referenced, the GED design of the essay scoring guide had to be descriptive, rather than prescriptive, in nature.
To define the scoring guide used for the evaluation of GED candidate

writing, GEDTS procured a large national stratified random sample of direct writing from actual graduating seniors. Because GEDTS had decided to use a six-point scale, the Writing Committee was asked to rank order the essays in six stacks based on the quality of the writing, and then to describe the general characteristics of the essays in each of the six stacks. Appendix A (excerpted from GEDTS, 1985b) further describes how the scoring guide was inductively developed from characteristics of high school senior essays written on one "anchor" topic. As Appendix A points out, the decision to apply one scoring guide across all future topics and readings helps to provide a foundation for scale stability across topics, readers, and scoring sessions. The scoring guide that was developed is given as Appendix C. The scoring standards as identified by the Committee in the scoring guide are not prompt-specific; the scoring guide, based on the description of the actual graduating high school students' essays, is intended to apply across topics. (Part of the topic selection process determines if field test topics result in essays that can appropriately be scored using the scoring guide.) These are invariant standards that remain in place over the years until a new norming is conducted.

Evaluating and Selecting Uniform Topics

To help minimize the problem of controlling topic variability, it was decided to restrict GED essay topics to one type (i.e., expository). An example of a topic is given in Appendix D (excerpted from GEDTS, 1985c). Judgmental criteria were developed to help the GED Writing Committee evaluate potential topics. Appendix E (excerpted from GEDTS, 1986) enumerates many of these judgmental criteria.

After potential topics are reviewed and edited by the Committee, they are field tested in administrations to both high school seniors and GED examinees. Papers written on the potential topics from these administrations are evaluated by members of the Writing Committee in a topic selection reading. The topic selection reading focuses on whether, in light of actual papers, the topics meet the criteria presented in Appendix E.

If topics "survive" these judgmental evaluations, statistical evaluations are performed. For high school seniors' scores on the set of papers, a Kruskal-Wallis test of the homogeneity of score distributions and a test of the equality of correlations between multiple-choice and essay scores are performed. To support the administration of any operational essay topic in combination with any multiple-choice test form, only those topics that yield non-significant results on both statistical tests are retained for operational use.

Selecting and Training Chief Readers and Readers

Because essays are scored in decentralized scoring sites and sessions, GEDTS determined it imperative to have a GED Testing Service trained and certified Chief Reader at each scoring site. This person is directly trained by the GEDTS Writing Assessment Specialist and is responsible for the recruitment and training of readers, the certification of the readers and the scoring site, and the operation of the independent scoring site. Each jurisdiction's GED Administrator is responsible for selecting a potential Chief Reader based on GED requirements. GED Chief Reader qualifications and certification requirements are included in Appendix F.

GEDTS decided to conduct all Chief Reader training in a central location (Washington D.C.) to help enforce the concept that all scoring sites are part of one program and that all readers must apply GED scoring standards uniformly. To eliminate as many variables as possible, all Chief Readers are trained using the same materials by the same trainer.

The training involves a two-and-a-half-day modified holistic training and qualifying session conducted by the GED Writing Assessment Specialist with some assistance from the GED Writing Committee. The major role of the Writing Assessment Specialist in training the Chief Readers, and the Chief Reader in training readers, is to communicate accurately the intent of the GED Official Scoring Guide and to focus discussion on attaining an understanding of the rationale of the Committee assigned scores as on the training essays related to the Scoring Guide.

The training consists of an orientation to the background and purpose of the assessment, a thorough review of the training manual, a discussion of reader objectivity issues (Appendix G), a review of guidelines for holistic scoring (Appendix H), exposure and participation in the Writing Committees' standard setting procedures, practice in scoring papers sets, and an opportunity to certify as a Chief Reader.

Periodic Writing Committee meetings are held to identify training materials. Members are first reoriented to the GED Official Scoring Guide through the scoring of a recalibration set to which they previously assigned scores. Committee members are then instructed to score essays based on the descriptions included in the Scoring Guide. This is necessary because even the Writing Committee, with 100 percent agreement, can drift from the score points as defined by the Scoring Guide. It is only by relating the selection of the score to the description of that score point in the Scoring Guide that this can effectively be detected.

The training sets--five sets of ten essays, and the certification sets-two sets of 20 essays, are standard essays as identified by the Writing Committee. This task is accomplished by making enough copies of a high school student writing sample to allow the Committee members to simultaneously read and discuss each essay. Committee members independently score each essay; the score is then compared to other Committee member scores. An essay is assigned a Committee or standard

score and included in the training set only if at least 80% of the members agree on the same score. In addition, Committee commentaries about why each essay receives a particular score are included to serve as rationale statements during the training.

The training sets are then assembled by the Writing Assessment Specialist, not as random sets with equal distribution at score points, but as systematic sets with a disproportionate number of borderline essays at each score point to train the readers to correctly discriminate between scoring points. Because the mean score on the GED essay for high school seniors is 3.67, and the mean score for GED candidates is 2.75, the training sets consist of a greater number of essays assigned scores of 2, 3, and 4. In addition, the training sets include examples of the problematic papers that cause extensive deliberation by the Committee and would be the probable cause of most discrepant scores requiring third readings (i.e., rap songs and abstract ideas.)

Using the initial training set, potential Chief Readers and readers are guided through a simulated Writing Committee essay selection exercise conducted to allow the participants to articulate score selection in the "language" of the Scoring Guide. This exercise serves two purposes. First, it shows the potential readers the procedure involved in assigning essay scores and emphasizes the importance of using the Scoring Guide in the discussion. Second, it allows the trainer to hear the rationale a potential reader is using when assigning a score to an essay. This can be beneficial to a Chief Reader who must select an alternate Chief Reader.

To qualify as a reader and as a Chief Reader, each candidate must demonstrate the ability to apply the GED Essay Scoring Guide standards by scoring reader certification sets with a minimum of 50 percent exact agreement and 50 percent contiguous agreement with pre-assigned Committee scores. If the potential Chief Reader or reader does not pass the first certifying round, the trainer provides additional training based on diagnosed weaknesses. The candidate is then given a second opportunity to certify.

The Chief Reader training session serves as a model for the Chief Readers to use when training potential readers for their scoring sites. After Chief Readers have been trained and certified by the GED Writing Assessment Specialist in Washington D. C., they are then responsible for recruiting (using GEDTS minimum reader qualifications and requirements specified in Appendix F), training (using the same procedures and materials presented in the Chief Reader training session), and certifying (using the same methods, materials and criteria used for Chief Readers) site readers. This is done to eliminate as many variables as possible between training and certifying Chief Readers and site readers.

Several GED Administrators attended the Chief Reader certification sessions so that they would be aware of the new program being implemented in their jurisdictions. It is interesting to note that some Administrators initially elected to have as many scoring sites as they

had Testing Centers. After attending the training sessions and learning of all the stringent requirements and certification procedures, ninety percent of the Administrators elected to operate only one scoring site.

Site Certification

After the Chief Reader and all readers for a particular scoring site have been certified, GEDTS then requires that the site be certified. Each scoring site must demonstrate the ability to score essays in accord with the standards defined by the GED Testing Service. To qualify for certification, a site must demonstrate the quality of its essay scoring in two respects:

1) reading reliability - the consistency with which different readers in a given reading award the same scores to a given set of essays, and
2) scale stability - the degree to which readers uniformly apply the GEDTS score scale (the scoring guide) in evaluating essays both within and across scoring sessions. This is determined by matching the total score awarded an essay by the potential site with the total score awarded an essay by GEDTS Writing Committee.

All essays included as site certification essays are original essays written by high school seniors or GED examinees. Only essays on which there was at least 80 percent agreement among the Writing Committee members are included in the certification sets. The majority of the essays have 100 percent agreement. There are thirteen sets of twenty essays, all written to the same topic; each set has an equal distribution among score points and Writing Committee agreement to limit variability of essays across certification sets.

Certification sets are assigned to readers by the GED Writing Assessment Specialist; each reader must read two sets of twenty essays and each certification set must be read twice. If there are any discrepant scores, the Chief Reader must read the essays and assign a third score. Chief Readers and readers do not know the GED Writing Committee scores for the site certification essays.

Site certification paper sets and scoring rosters are then returned to GEDTS for evaluation. Appendix I shows a sample table that illustrates the results reported back to a scoring site after a site certification or monitoring reading. After discussing the results and the implications of the results of the initial jurisdictions who completed site certification, GEDTS established the following site certification standards:

PERCENTAGE OF AGREEMENT WITH WRITING COMMITTEE TOTAL ESSAY SCORES*	STATUS	IMPLICATIONS
90-100%	Pass	The site meets the standards as established by GEDTS and may begin scoring essays.
85-90%	Provisional Pass	The site conditionally meets the standards as established by GEDTS and may begin scoring essays with the understanding that additional training and a second certification must occur within 60 days after receipt from GEDTS of a diagnostic report of the scoring site and readers.
84% and below	Failure	The site does not meet the standards as established by GEDTS and may not score essays until a second certification reading is conducted and a passing score is obtained. GEDTS will send a diagnostic report identifying the problem areas and/or readers to help prepare for the second session.

* AGREEMENT OF THE SCORING SITE AND THE WRITING COMMITTEE TOTAL ESSAY SCORES IS DEFINED AS SITE SCORES WITHIN TWO POINTS OF THE WRITING COMMITTEE SCORES.

The major concern of GED is the agreement of the reader with the GED Writing Committee score. It is possible for sites to have a very high agreement between readers ("reading reliability") but a very low agreement with the Writing Committee scores. Because scoring is decentralized, it is critical to relate the reader and site scores with the GED Writing Committee scores to prevent a drift to local or individual scoring standards. A diagnostic report evaluating individual reader as well as site performance is returned to each Chief Reader. In the final analysis, fifty-five jurisdictions have only one scoring site, while five jurisdictions have multiple scoring sites: one jurisdiction has two sites, two jurisdictions have three sites, one

jurisdiction has five sites, and one jurisdiction has seven sites. It is of interest that both the jurisdiction with five sites and the jurisdiction with seven sites are seriously considering consolidating sites to make the operation more manageable.

Twenty-six jurisdictions have one central scoring site, fifteen are using the GED Essay Scoring Service, eleven are using the Iowa Consortium (three scoring sites located in Iowa and coordinated by the Iowa GED Administrator), and eight jurisdictions are using commercial scoring services.

Site Monitoring

Another major task in operating a program that allows decentralized scoring is the requirement to reduce or optimally eliminate scoring drift. Kadzin (1977) defined consensual observer drift as "the tendency of observers to change the manner in which they apply the definitions of behavior over time." Because this can impact scoring accuracy or scale stability without affecting inter-rater reliability, GED instituted three monitoring strategies to ensure score scale stability and eliminate scale drift: Chief Reader monitoring, monitoring through recalibration readings, and site monitoring.

Because GEDTS can monitor scoring sites on only a periodic basis, the Chief Reader's monitoring role is critical. The Chief Reader can check each reader's scoring from session to session, thus discerning if readers consistently adhere to the GEDTS scoring standard. By also serving as a reader, the Chief Reader is required to complete second readings after as many readers as possible during each scoring session. Even though a reader may always award a contiguous score, the Chief Reader, when serving as a second reader, may identify a reader's tendency to consistently score papers one point higher or lower than the expected score as dictated by the GED Official Scoring Guide. If the Chief Reader identifies a pattern for a particular reader, the Chief Reader must then work with that reader to help bring the reader back on scale. In addition, since the Chief Reader resolves all discrepant scores, the Chief Reader can record each time a reader assigns a discrepant score and notify the reader of the scoring drift, thereby helping the reader get on scale.

As a second method of measuring score scale stability on a scoring session basis, GEDTS provides recalibration sets. At the beginning of each scoring session and at the introduction of a new topic, all readers are required to score a set of five recalibration essays that were read and assigned scores by the Writing Committee. Readers compare their scores with the Writing Committee scores and check to see if they have remained consistent with the GED Scoring Guide standards. Readers also check to determine if they are rating recalibration sets consistently too high or too low and need to review the Scoring Guide. Therefore, not only reader accuracy to the score point but reader drift in either a high or low direction is monitored by using the recalibration paper sets. The GED Writing Committee has developed a minimum of thirteen recalibration sets for each operational topic.

Finally, GEDTS will conduct controlled site monitoring on a periodic basis in two ways: systematic monitoring and random monitoring. Both of these procedures are based on the same standard--scores awarded by the Writing Committee.

Systematic Site Monitoring

Systematic monitoring is essentially the same as the site certification procedure with the same guidelines and criteria for passing. The monitoring essays are identified by the Writing Committee and monitoring essay sets are developed by the GED Writing Assessment Specialist using the same standards and procedures as were followed with the certification sets. All certified scoring sites are simultaneously mailed four monitoring sets of ten essays and instructed to include the monitoring sets in the next scheduled scoring session. Every reader is required to read all four sets of ten essays. All readers who were certified after the certification of the site are required to participate in site monitoring. Site Chief Readers and readers do not know the standard scores given by the Writing Committee to the essays in the monitoring sets.

Scoring rosters and reader identification sheets are returned to GEDTS. GEDTS then compares the scores of the individual readers and the scoring sites against the scores assigned by the Writing Committee. Again, it is essential that the readers agree with the Writing Committee scores. This is a quality check to compare the scoring sites score scale stability with the GED scoring standards. Only then can GED Testing Service conclude that a scoring site has maintained scale stability. Site monitoring reports are identical to site certification reports. Copies are mailed to the Chief Reader for diagnostic and evaluation purposes.

In addition, because all readers in all sites are reading and scoring the same forty essays, GEDTS is able to evaluate score scale stability across scoring sites in an attempt to assure that an essay awarded a total score in any one of the sites will receive a very similar score at any other site. It is argued that systematic monitoring may produce the Hawthorne effect because readers are aware that their reading is being monitored and therefore may yield greater than usual agreement with the Writing Committee scores. First, the reader has to understand and accept the Scoring Guide to be able to knowingly apply the criterion when scoring essays. Second, some readers' anxiety may cause them to over-analyze the essays and thus have weaker agreement. These suppositions will be further addressed in the major findings section of this paper.

When evaluating readers performance on the same forty essays, GEDTS is able to collapse all scoring sites into one large site and evaluate inter-rater reliability. This is done by tallying all readers scores to each of the forty essays and assessing what percent of the readings resulted in the same scores, contiguous scores (scores within one point in either direction), and discrepant scores (reader scores differing by more than one point). This allows GEDTS to check the reading reliability

of sites and to identify sites with a higher than average percent of discrepant scores and lower than average reading reliability. These results are reported to the chief readers as diagnostic information on the sites performance. In addition, GEDTS may use this type of data to identify problematic essays that require more attention in preparation of future recalibration materials.

Random Site Monitoring

The second type of controlled monitoring, random monitoring, is possible because GEDTS is embedding field test items in the operational test forms and requiring Testing Centers to return their answer sheets to GEDTS for scanning and item evaluation. Because of this, the GED Writing Assessment Specialist has access to essays scored at a number of certified scoring sites.

For this monitoring, a random sample of forty scored essays is selected from each certified scoring site that returns essays to GEDTS. The decision to select forty essays was made to duplicate the number of essays in both the site certification and systematic monitoring procedures. Scores recorded on the essays by the readers at the scoring site are masked and then copies are made of the essays. Writing Committee members read each essay simultaneously and assign independent scores. Scores are discussed, tallied, and a total score is recorded for each essay. Scoring site total scores are compared with the Writing Committee total scores to evaluate scoring accuracy. This procedure differs from the systematic monitoring in two significant respects. First, readers do not know that their scores on these particular essays will be evaluated, and second, the essays (and topics) vary across sites. The constant or standard in both monitoring procedures is the Writing Committee scores.

Random monitoring provides an additional measure of score scale stability for both the scoring site and the Writing Committee. For both site certification and systematic monitoring, the mean essay score for a given set of essays is predetermined by the composition of the essays as assembled by the GEDTS Writing Assessment Specialist. In addition, the systematic monitoring essay sets only contain papers with even scores (i.e., 2, 4, 6, 8, 10, and 12) because only those essays that had eighty to one hundred percent Writing Committee score agreement are used. In contrast, in random monitoring, the mean essay score is determined by the random sample of essays from a particular site. If the mean essay score for a scoring site is significantly lower or higher than the GED mean essay score, the Writing Committee can serve as a "check." The GEDTS (Writing Committee) mean essay scores should vary from site to site contingent on the quality of writing in each sample of essays from the particular scoring sites and be very near the mean score assigned by the scoring site. This comparison of systematic and random monitorings is addressed more fully in the major findings section of this paper.

Because some scoring sites return data tapes rather than answer sheets, it is currently not possible to have random monitoring results for all

scoring sites. However, the sample of sites evaluated and reported in this paper is thought to be representative of all scoring sites.

Major Findings

Chief Reader Training

Approximately 150 persons attended the six Chief Reader training sessions conducted by the GED Writing Assessment Specialist in Washington D. C. from July 1988 through February 1989. Only four participants needed to complete the second certification set. All other participants passed the initial certification set with a mean exact agreement of 65 percent and a mean contiguous agreement of 35 percent. This success can be contributed to the initial Chief Reader qualification requirements, the careful screening and selection of potential Chief Reader candidates, and the design and development of the training materials and the total training program.

Site Certification

At the end of the first year of site certification, there are fifty-two operational GED scoring sites. Seventy-seven percent of the scoring sites passed on the first certification attempt with a mean score of ninety-four percent agreement. Twenty-three percent of the sites received a conditional pass on the first attempt with a mean score of eighty-seven percent agreement, and then passed a second certification session with a mean score of ninety-one percent agreement. Only three potential scoring sites failed site certification; all three of these sites were in a jurisdiction that initially elected to have eleven sites. When the sites failed, the GED Administrator elected to consolidate and only operate seven sites. It should also be noted that in the case of the three failures and in most instances where a second site certification was required, the training was conducted by an alternate Chief Reader or site Chief Reader—trainers who were once removed from the central training in Washington, D. C.; more specifically, those jurisdictions where the Chief Reader trained readers at multiple sites and appointed an alternate Chief Reader to be site leader who then conducted the certification. This data indicates that GEDTS must require each scoring site to have a Chief Reader directly trained by GEDTS in Washington D.C. This finding may further influence the two jurisdictions with multiple scoring sites to consolidate sites.

Site Monitoring

Monitoring results of fifteen sites are presented in this paper. Tables 1a through 1d show: the random monitoring results, the systematic monitoring results, a comparison of random and systematic monitoring results for two scoring sites, and a comparison of the systematic monitoring results with the original certification results for ten of the

scoring sites participating in the systematic monitoring. Each table shows the degree of agreement between the site scoring and the Writing Committee scoring of the monitored essay papers. In addition to showing the percentage of papers scored the same, one point differently, two points differently, and more than two points differently, the tables show the intraclass correlation of the sites' scores and the Writing Committee's scores. (See the columns labelled h in Tables 1a through 1d.) The intraclass correlations were calculated as described by Ebel (1967). The "between rater" variance is included in the error terms for calculations in order to ensure that differences in applying the scoring standards between the site and Writing Committee are reflected in the intraclass correlations. The intraclass correlation is the preferred coefficient for reflecting the agreement between the site and Writing Committee and the agreement between readers within the site on given papers because it is sensitive to shifts (scale drift) in the application of the scoring standards. Correlation coefficients such as the Pearson are not sensitive to such shifts.

Intraclass correlations are also calculated for readers within scoring site. Application of the Spearman Brown formula to the intraclass correlation for readers within site results in a measure of the sites "reading reliability." While of interest, it is important to note that reading reliability in and of itself provides no assurance of scale stability. Scale stability in this setting is addressed by the agreement of the site scores and the Writing Committee scores on the same papers.

As indicated by the tables, all but one of the 15 monitored scoring sites unconditionally passed the random and/or the systematic monitoring using the criteria described earlier in the discussion of site certification. That is, only one site had greater than ten percent of their essay total scores (on the 2 to 12 scale) differing by more than two points from the Writing Committee scores. By this criteria, the data show that the efforts to maintain score scale stability in decentralized scoring sites seem to be working well.

(INSERT TABLE 1a ABOUT HERE)

Table 1a reports results for the six sites participating to date in the random monitoring. To review, the random monitoring essays are actual GED examinees' essays which were read and scored during normal scoring sessions by the readers at the scoring sites. Based on the sample of six scoring sites represented in the Table 1a, the percentages of "agreement" between essay scores from the scoring sites and from the GEDTS Writing Committee were: 56.7 percent of scores were equal; 33.8 percent of scores differed by plus or minus one point; and 7.7 percent of scores differed by plus or minus two points on the twelve point scale. Only 1.7 percent of the essays received scores from the site and the Writing Committee that differed by more than two points.

It is interesting to note in Table 1a that means and standard deviations of the Writing Committee and the scoring sites vary from site to site. These differing distributions of scores by scoring sites and by the

Writing Committee reflect that the level of the examinee performance on the essay (and the total GED Test battery) varies from site to site. The data in Table 1a tends to verify that it is possible to maintain score scale stability across multiple sites even when the distributions of scores across site vary. This evidence also substantiates the Writing Committee's and the scoring sites' ability to apply the Official Scoring Guide scale in different sites with different distributions of scores and still remain true to Official GED Essay Scoring Guide scale.

Comparisons of the intraclass correlations across the sites shown in Table 1a (columns h and n) must be made cautiously due to the differing variances of essay scores across sites. Like other correlation coefficients, the intraclass coefficients tend to be lower when applied to distributions with small variances. This possibly explains the seemingly very low "reading reliability" for site E. Site E had the lowest variation of scores within its sample of essays. Comparisons of intraclass values in columns h and n across Tables 1a, 1b, and 1c should also be made cautiously because of the differences in variance of the essay score distributions. A potentially helpful procedure would be to adjust the intraclass correlations to reflect what they would have been had all random samples of essays drawn for the random monitoring had the same variance. The Writing Committee overall variance of the randomly drawn essays would serve as an appropriate "standard variance" if comparisons were to be made within Table 1a. Or, if comparisons are desired across Tables 1a and 1b, the Writing Committee's variance of the systematic monitoring sets of essays could be used as a basis for adjusting the intraclass values obtained for the random monitoring. In this regard, the systematic monitoring has an advantage over the random monitoring because all monitored sites read the same selected set of essays. That is, because the variance of site scores for the systematic essays across sites should be equal given that all sites read the same papers, comparisons of the intraclass values across sites participating in the systematic monitoring can be made without regard for differences in the distributions of scores for the samples of essays read.

(INSERT TABLE 1b ABOUT HERE)

Table 1b shows the results of the systematic monitoring. The systematic monitoring essays are essays scored by the Writing Committee then compiled into sets by the GEDTS Writing Assessment Specialist. The mean score and standard deviation for these essay sets remains constant (i.e., 6.40 and 2.02 respectively though slight differences from 2.02 result from differences in the number of total essay scores produced by the number of readers at the site). These monitoring sets are then mailed to the scoring sites and included in the next normal scoring sessions by the readers at the scoring sites. Based on the sample of eleven scoring sites represented in Table 1b, the percentages of "agreement" between scores from the scoring sites and from the GEDTS Writing Committee were: 41.2 percent of scores were equal; 36 percent of scores differed by plus or minus one point; and 17.8 percent of scores differed by plus or minus two points on the twelve point scale. Only five percent of the essays received scores from the site and the Writing Committee that differed by

more than two points.

Although the difference in mean scores between the scoring sites and the GEDTS Writing Committee is similar on both the random and systematic monitoring sets, the performance on the systematic monitoring differs from that on the random in one significant way. Comparison of columns d and f means in Tables 1a and 1b shows that scoring sites awarded approximately sixteen percent fewer equal scores and about ten percent more scores differing by two points in the systematic monitoring than in the random monitoring. These results are further illustrated in Table 1c that shows random and systematic monitoring results for two scoring sites that participated in both the random and the systematic monitoring.

(INSERT TABLE 1c ABOUT HERE)

There are two probable causes for this behavior; the composition of the essay sets and the effects of readers knowing they are being evaluated. First, the composition of the systematic monitoring sets differs from the random sets in that the systematic sets have only essays that the Writing Committee awarded the same score - thus the sets have only even-numbered scores on the two to twelve scale. There are only six GEDTS Writing Committee total scores (i.e., 2, 4, 6, 8, 10, and 12). In the random monitoring sets, the Writing Committee may have split scores (i.e., fifty percent assigned a score of three and fifty percent assigned a score of four resulting in a score of seven.) Thus, in the random monitoring essay sets, there are ten scores, both even and odd, that can be awarded (i.e., 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12). Also, because the random essays are essays from actual scoring site readings, they are more characteristic of essays the scoring site readers encounter on a regular basis. Conversely, the systematic essay sets include essays from the entire range of two to twelve, presenting readers with better essays than usually read, thus increasing the difficulty of assigning an exact score.

Second, randomly monitored readers do not know that scores awarded to essays during actual scoring sessions will ever be evaluated. Because of this, the readers evaluated by random monitoring essays are more likely to have quickly read and scored the essays as taught during training. During the systematic monitoring session, the readers are aware of which essays scores are for monitoring purposes and that their scores will be sent to GEDTS for evaluation. This can cause reader anxiety or cause the reader to over-analyze the essays before awarding scores.

(INSERT TABLE 1d ABOUT HERE)

Table 1d compares the original certification and subsequent monitoring results some months to a year later. Table 1d column means suggest that the sites had a higher exact agreement and fewer discrepant scores on the systematic monitoring than they had on the original certification. A plausible explanation for this improvement includes both the constant vigilance of the Chief Reader as monitor on a per session basis, and the required recalibration readings to Writing Committee scored essay sets before each scoring session and each new topic. These steps are both

essential and critical to the success of the program.

After reviewing and comparing the first results of both the random monitoring in Table 1a and the systematic monitoring in Table 1b, the random monitoring appears to be the more realistic of the two systems. Because the random monitoring is neither a synthetic monitoring system nor a system composed of only essays with eighty to one hundred percent Writing Committee agreement, the scores of the readers and the Writing Committee tend to agree more than is the case with scores of the readers and the Writing Committee on the systematic essay sets. However, the systematic monitoring does provide checks that cannot be accomplished with random monitoring. For example, systematic monitoring allows evaluation of site scoring throughout the range of possible scores. Also, since random monitoring essays differ across sites, it is not possible to evaluate reader performance by essay-score-point-by-site or to collapse all scoring sites into one universal site as is possible with systematic monitoring. These analyses for systematic monitoring essay sets are illustrated below in Tables 2a through 2c.

(INSERT TABLE 2a HERE)

Table 2a shows reader agreement by essay score point on the one to six reader score scale for the systematic monitoring essays across all scoring sites. The universal problem illustrated in Table 2a is with the four and five score points. In the instance of both the four and five score points, less than fifty percent of the total population of readers awarded the same score as the Writing Committee. Additionally, in both instances, the readers awarded a score lower than the Writing Committee's score. There are plausible explanations for this difference: readers had less training at these score points, readers see fewer four and five papers during actual scoring sessions and are reticent to award these scores, or the language of the Official GED Essay Scoring guide is not as clear at these score points.

Because a) the mean essay score for GED examinees is 2.97 and fewer GED examinee essays are of the quality of the four and five score points, and b) the GED Tests cut score is at about the 33rd percentile of the weighted composite score scale, the difficulty at these score points is not quite as critical as it would be at the two or three score points. However, immediate steps to improve accuracy at the four and five score points include: a) providing more four and five essays in the recalibration sets for each topic, b) informing the Chief Readers of the reported difficulty with these score points and, c) instructing Chief Readers to encourage attention and clarification at these score points during training and scoring.

(INSERT 2b and 2c ABOUT HERE)

Looking at Tables 2b and 2c against Table 2a provides comparisons of the scoring by Sites J and L on the systematic monitoring essays by score point against the total sample of systematic monitoring readers. By comparing a particular scoring site with the total sample of readers

from systematic monitoring sites, site idiosyncracies can be diagnosed.

The systematic monitoring data for scoring site J reported in Table 2b is similar to that of the data for the total sample of systematic readers reported in Table 2a. The problematic score points are identical; that is, the four and the five score points. Site J, however, does show larger discrepant score percentages at these score points, so the Chief Reader will be told to work specifically on four and five papers.

Table 2c shows the site L scores by score point. Site L was the only site failing to meet the unconditional pass criteria for certification and monitoring. The site L systematic monitoring results shown in Table 2c reveal a different problem. Scoring site L shows a mean agreement of less than fifty percent at the three score point. In addition, the readers' awarded a score higher than the Writing Committee score at the two, three, and four score scale points. This tendency to score high is supported by the difference in mean scores for Site L's systematic monitoring shown in Table 1b. The mean score for the Writing Committee was 6.40, while site L's mean score for the same essays was 7.15. Because the site must attempt a recertification, the Chief Reader will be assisted in interpreting and acting on the diagnostic report that in addition to the general report (shown as Appendix I for site L) will include Table 2c. The site L information contained in Appendix I and Table 2c (in addition to an analysis by score point like that shown in Table 2c for each reader) highlights the need for additional attention at site L to the two, three, and four score scale points. It should be noted that the GEDTS Writing Assessment Specialist monitors sites such as site L more frequently than those sites which yield analyses by score point more closely parallel to that of the Writing Committee scores.

Finally, the data presented in Tables 1a through 1d and Table 2a assist in the consideration of different certification criteria. For example, an alternative criteria might be the evaluation of the percentage of readings differing from the Writing Committee scores by 2 or more points on the six point scale. Table 2a shows the following percentages of such discrepant readings: zero percent at score point six; eight percent at score point five; four percent at score point four; four percent at score point three; one percent at score point two; and zero percent at score point one. These findings indicate that the criteria might reasonably be changed to ninety-five percent agreement with the Writing Committee score on the one to six point scale for an unconditional pass, ninety to ninety-four for a conditional pass, and eighty-nine and below as failing scores. Implementing this more rigorous certification criteria would further improve the extent to which the GEDTS program enforces the required consistency of essay scoring across sites.

Another consideration relating to the criteria for certification is the use of the intraclass correlation between the scoring site and the Writing Committee total scores. The advantage of the intraclass correlation over the percentage agreement within 2 points of the Writing Committee scores on the total score scale is that the intraclass is sensitive to differences across sites in the percentage of scores awarded

that are equal, one different, two points different, etc. In essence, the intraclass is sensitive to scale drift of the site readers from the scoring standards defined by the Writing Committee on the original "norm group" of high school senior essays. The data in Tables 1a through 1d suggest that intraclass correlations of the site and Writing Committee total scores below about 0.80 warrant a close look at the site's performance. An additional motive for using the intraclass correlation as an idicator of the criteria of the site certification and monitoring. is that the reliability of the weighted composite scores of the essay and multiple-choice parts of the Writing Skills Test needs to be sufficiently high to ensure accuracy near the cut score. (For a description of the formation of the GED Writing Skills weighted composite score, see Patience and Swartz, 1987.)

Conclusions and Recommendations

When a direct assessment of writing is used, it is mandatory to establish and implement scoring standards and to initiate training, certification, and monitoring procedures which provide continuity among scoring sites and scoring sessions. The scoring process for all GED scoring sites is continually evaluated: both on a regular basis by the Chief Reader and systematically by the GED Testing Service. Ensuring fair and stable scoring from scoring site to scoring site and reading session to reading session requires careful planning and systematic procedures. These conclusions are applicable to both small programs and much more complex programs such as the decentralized GED Testing Service situation.

Will the process described herein result in acceptably reliable GED Writing Skills scores? What has been learned about the effectiveness of these procedures, and what should others know when designing such a program?

Some judgments gained thus far from the experience are clear. The purpose for the direct assessment must be identified and articulated so that an integrated model can be designed from the onset--form follows function. A committee should be established to give "face validity" to the program and to provide a historical perspective and continuity from administration to administration. This committee should be involved in all critical stages of program development and evaluation (i.e., developing the scoring guide, selecting topics, identifying training essays, etc.) Development of the standards, training, and certifying materials is a meticulous process which establishes the success or failure of the program. There are no discrepant papers, only discrepant readings. Successful internalization of the score scale by readers is contingent on the clarity and effectiveness of the reader training and monitoring. Chief Readers at each site must be first generation - that is, trained by the source. Chief Readers must be able to understand and accept the criteria and must have the skills necessary to articulate the scoring standards to other readers. Readers must be sufficiently trained and then recalibrated on a per scoring session per topic basis. Chief Readers and readers must so internalize the scoring standards as to

become as objective as computers or robots programmed to run only one program. Nothing can be left to chance. Constant vigilance is required to maximize essay score scale stability and reading reliability. This vigilance is essential given that achieving acceptable test-retest score reliability depends, in large measure, on first establishing essay scoring scale stability and reading reliability.

Table 1a
Random Monitoring Score Scale Stability Results

Site/ Mon. Type	Number of Readers	Number of Essays/ N-scores	Agreement of Scoring Site vs. Writing Comm. Scores: % Equal	% 1pt. diff	% 2pt. diff	% >2pt. diff	intra class corr.	Essay Score Means & Stan. Dev.: Writ. Comm. $\bar{x}$	Writ. Comm. sd	Site $\bar{x}$	Site sd	Reading Reliability: % disc.	S.B. rel.
a	b	c	d	e	f	g	h	i	j	k	l	m	n
A/r	5	38/38	52.6	31.6	15.8	0.0	0.75	5.13	1.55	5.39	1.20	0.0	0.80
B/r	2	39/38	64.1	33.3	2.6	0.0	0.93	6.36	1.75	6.23	1.81	0.0	0.88
C/r	3	39/39	76.9	15.4	7.7	0.0	0.84	6.08	1.24	6.08	1.18	0.0	0.76
D/r	4	40/40	67.5	17.5	15.0	0.0	0.76	6.68	1.29	6.30	1.24	0.0	1.00
E/r	9	39/39	28.2	61.5	2.6	7.7	0.60	6.38	1.48	5.64	1.09	0.0	0.44
F/r	4	39/39	51.3	43.6	2.6	2.6	0.60	6.38	1.16	6.46	1.54	0.0	0.85
Column Means	5	39/39	56.7	33.8	7.7	1.7	0.75	6.17	1.41	6.01	1.34	0.0	0.79

Notes. Column a: Shows site identification and type of monitoring (r=random; s=systematic).
Column g: Shows percent of site's essay scores (on 2 to 12 scale) that differed by more than 2 points. from the GED Writing Committee Scores. To be certified unconditionally, a site must have no more than 10% of its papers scored 2 or more points differently than the GED Writing Committee scores.
Column h: Shows the intraclass correlation of the writing committee and scoring site total essay scores (on the 2 to 12 essay total score scale).
Column m: Shows the percent of monitoring essays with two readers' scores differing by 2 or more points. (Each reader's score is on the 1 to 6 scale.) Such readings are deemed discrepant and require a third reading by the Chief Reader.
Column n: Shows the "reading reliability" of the site's scoring of the monitoring essays. This is calculated by applying the Spearman-Brown Formula to the intraclass correlation of the site's 1st, 2nd, and 3rd (when present) reader scores (on the 1 to 6 scale).

Table 1b
Systematic Monitoring Score Scale Stability Results

Site/ Mon. Type	Number of Readers	Number of Essays/ N-scores	Agreement of Scoring Site vs. Writing Comm. Scores: % Equal	% 1pt. diff	% 2pt. diff	% >2pt. diff	intra class corr.	Essay Score Means & Stan. Dev.: Writ. Comm. $\bar{X}$	Writ. Comm. sd	Site $\bar{X}$	Site sd	Reading Reliability: % disc.	S.B. rel.
a	b	c	d	e	f	g	h	i	j	k	l	m	n
C/s	17	40/360	42.2	41.1	16.1	0.6	0.85	6.40	2.01	6.26	1.89	4.2	0.85
F/s	8	40/160	37.5	43.8	13.1	5.6	0.81	6.40	2.02	6.69	1.95	11.3	0.80
G/s	14	40/280	40.0	44.3	12.1	3.6	0.85	6.40	2.01	6.61	2.07	13.6	0.79
H/s	4	40/80	43.8	31.3	22.5	2.5	0.82	6.40	2.02	6.49	2.00	8.8	0.85
I/s	5	40/120	43.3	43.3	10.8	2.5	0.85	6.40	2.02	6.25	1.92	5.0	0.82
J/s	3	40/80	38.8	36.3	15.0	10.0	0.78	6.40	2.02	6.21	2.12	2.5	0.88
K/s	5	40/120	40.0	38.3	18.3	3.3	0.83	6.40	2.02	6.43	2.13	10.8	0.79
L/s	3	40/80	15.0	53.8	18.8	12.5	0.72	6.40	2.02	7.15	2.16	7.5	0.82
M/s	3	40/80	66.3	3.8	27.5	2.5	0.82	6.40	2.02	6.44	2.12	0.0	0.99
N/s	3	40/80	42.5	28.8	21.3	7.5	0.77	6.40	2.02	5.91	2.09	0.0	0.93
O/s	6	40/120	44.2	30.8	20.0	5.0	0.82	6.40	2.02	6.24	2.14	0.8	0.91
Column Means	6	40/142	41.2	36.0	17.8	5.0	0.81	6.40	2.02	6.43	2.05	5.86	0.86

Notes. Column a: Shows site identification and type of monitoring (r=random; s=systematic).
Column g: Shows percent of site's essay scores (on 2 to 12 scale) that differed by more than 2 points. from the GED Writing Committee Scores. To be certified unconditionally, a site must have no more than 10% of its papers scored 2 or more points differently than the GED Writing Committee scores.
Column h: Shows the intraclass correlation of the writing committee and scoring site total essay scores (on the 2 to 12 essay total score scale).
Column m: Shows the percent of monitoring essays with two readers' scores differing by 2 or more points. (Each reader's score is on the 1 to 6 scale.) Such readings are deemed discrepant and require a third reading by the Chief Reader.
Column n: Shows the "reading reliability" of the site's scoring of the monitoring essays. This is calculated by applying the Spearman-Brown Formula to the intraclass correlation of the site's 1st, 2nd, and 3rd (when present) reader scores (on the 1 to 6 scale).

Table 1c
Comparison of Random and Systematic Monitoring
Score Scale Stability Results

Site/ Mon. Type	Number of Readers	Number of Essays/ N-scores	Agreement of Scoring Site vs. Writing Comm. Scores: % Equal	% 1pt. diff	% 2pt. diff	% >2pt. diff	intra class corr.	Essay Score Means & Stan. Dev.: Writ. Comm. $\bar{X}$	Writ. Comm. sd	Site $\bar{X}$	Site sd	Reading Reliability: % disc.	S.B. rel.
a	b	c	d	e	f	g	h	i	j	k	l	m	n
C/r	3	39/39	76.9	15.4	7.7	0.0	0.84	6.08	1.24	6.08	1.18	0.0	0.76
C/s	16	40/360	42.2	41.1	16.1	0.6	0.85	6.40	2.01	6.26	1.89	4.2	0.85
F/r	4	39	51.3	43.6	2.6	2.6	0.60	6.38	1.16	6.46	1.54	0.0	0.85
F/s	8	40/160	37.5	43.8	13.1	5.6	0.81	6.40	2.02	6.69	1.95	11.3	0.80
Random Column Means	3	39/39	64.1	29.5	5.2	1.3	0.72	6.23	1.20	6.27	1.36	0.0	0.81
Systematic Column Means	12	40/260	39.9	42.5	14.6	3.1	0.83	6.40	2.02	6.48	1.92	7.8	0.83

Notes.
Column a: Shows site identification and type of monitoring (r=random; s=systematic).
Column g: Shows percent of site's essay scores (on 2 to 12 scale) that differed by more than 2 points. from the GED Writing Committee Scores. To be certified unconditionally, a site must have no more than 10% of its papers scored 2 or more points differently than the GED Writing Committee scores.
Column h: Shows the intraclass correlation of the writing committee and scoring site total essay scores (on the 2 to 12 essay total score scale).
Column m: Shows the percent of monitoring essays with two readers' scores differing by 2 or more points. (Each reader's score is on the 1 to 6 scale.) Such readings are deemed discrepant and require a third reading by the Chief Reader.
Column n: Shows the "reading reliability" of the site's scoring of the monitoring essays. This is calculated by applying the Spearman-Brown Formula to the intraclass correlation of the site's 1st, 2nd, and 3rd (when present) reader scores (on the 1 to 6 scale).

Table 1d
Comparison of Systematic Monitoring and Systematic Certification Results

Site/ Mon. Type	Number of Readers	Number of Essays/ N-scores	Agreement of Scoring Site vs. Writing Comm. Scores: % Equal	% 1pt. diff	% 2pt. diff	% >2pt. diff	intra class corr.	Essay Score Means & Stan. Dev.: Writ. Comm. $\bar{X}$	Writ. Comm. sd	Site $\bar{X}$	Site sd	Reading Reliability: % disc.	S.B. rel.
a	b	c	d	e	f	g	h	i	j	k	l	m	n
F/s	8	40/160	37.5	43.8	13.1	5.6	0.81	6.40	2.02	6.69	1.95	11.3	0.80
F/c	8	140/140	27.9	47.1	19.3	5.7	0.71	5.79	1.87	5.49	2.05	9.3	0.81
G/s	14	40/280	40.0	44.3	12.1	3.6	0.85	6.40	2.01	6.61	2.07	13.6	0.79
G/c	6	100/99	30.3	50.5	15.2	4.0	0.79	5.92	1.96	6.16	2.00	6.1	0.79
H/s	4	40/80	43.8	31.3	22.5	2.5	0.82	6.40	2.02	6.49	2.00	8.8	0.85
H/c	6	120/120	26.7	45.8	20.7	6.7	0.77	5.88	1.93	6.02	2.20	15.0	0.72
I/s	5	40/120	43.3	43.3	10.8	2.5	0.85	6.40	2.02	6.25	1.92	5.0	0.82
I/c	24	260/260	25.8	55.0	12.7	6.5	0.77	5.85	1.90	5.88	2.07	10.4	0.78
J/s	3	40/80	38.8	36.3	15.0	10.0	0.78	6.40	2.02	6.21	2.12	2.5	0.88
J/c	4	80/80	27.5	48.8	13.8	10.0	0.53	5.73	1.82	5.43	1.87	8.8	0.75
K/s	5	40/120	40.0	38.3	18.3	3.3	0.83	6.40	2.02	6.43	2.13	10.8	0.79
K/c	7	140/139	33.8	41.0	20.1	5.0	0.76	5.91	1.95	5.52	1.95	7.9	0.81
L/s	3	40/80	15.0	53.8	18.8	12.5	0.72	6.40	2.02	7.15	2.16	7.5	0.82
L/c	13	260/260	33.8	44.2	16.9	5.0	0.77	5.85	1.90	5.89	1.90	13.1	0.72
M/s	3	40/80	66.3	3.8	27.5	2.5	0.82	6.40	2.02	6.44	2.12	0.0	0.99
M/c	3	40/39	35.9	28.2	28.2	7.7	0.67	5.95	1.81	6.59	2.16	0.0	0.94
N/s	3	40/80	42.5	28.8	21.3	7.5	0.77	6.40	2.02	5.91	2.09	0.0	0.93
N/c	3	40/40	35.0	47.5	17.5	0.0	0.84	6.00	1.97	5.93	1.89	10.0	0.77
O/s	6	40/120	44.2	30.8	20.0	5.0	0.82	6.40	2.02	6.24	2.14	0.8	0.91
O/c	6	100/99	21.2	47.5	20.2	11.1	0.63	5.84	1.89	5.11	1.82	13.1	0.66
Monitoring Means	6	40/120	41.1	35.5	17.9	5.5	0.81	6.40	2.02	6.44	2.07	6.03	0.86
Certification Means	7	128/128	29.8	55.6	18.5	6.2	0.72	5.87	1.89	5.80	1.99	9.37	0.78

Notes. Column a: Shows site identification and type of monitoring (s=systematic monitoring; c=systematic certification).

Column g: Shows percent of site's essay scores (on 2 to 12 scale) that differed by more than 2 points. from the GED Writing Committee Scores. To be certified unconditionally, a site must have no more than 10% of its papers scored 2 or more points differently than the GED Writing Committee scores.

Column h: Shows the intraclass correlation of the writing committee and scoring site total essay scores (on the 2 to 12 essay total score scale).

Column m: Shows the percent of certification/monitoring essays with two readers' scores differing by 2 or more points. (Each reader's score is on the 1 to 6 scale.) Such readings are deemed discrepant and require a third reading by the Chief Reader.

Column n: Shows the "reading reliability" of the site's scoring of the certification/monitoring essays. This is calculated by applying the Spearman-Brown Formula to the intraclass correlation of the site's 1st, 2nd, and 3rd (when present) reader scores (on the 1 to 6 scale).

Table 2a.

Systematic Monitoring
Total Reader Population

Performance by Score Point

			Reader Agreement with GED Essay Score				
			- 2 Points	- 1 Point	Same Score	+ 1 Point	+ 2 Points
			Percent	Percent	Percent	Percent	Percent
Writing Score	Committee Essays	Readings					
06	1	71	.	15.49	84.51	.	.
05	3	213	7.98	38.50	41.31	12.21	.
04	9	639	2.19	37.40	43.04	15.65	1.72
03	18	1278	0.31	16.74	59.55	19.64	3.76
02	8	568	.	6.51	68.13	24.47	0.88
01	1	71	.	.	83.10	16.90	.

Table 2b

Site J Systematic Monitoring
Total Reader Population

Performance by Score Point

			Reader Agreement with GED Essay Score				
			- 2 Points	- 1 Point	Same Score	+ 1 Point	+ 2 Points
			Percent	Percent	Percent	Percent	Percent
Score	Essays	Readings					
06	1	3	.	.	100.00	.	.
05	3	9	11.11	55.56	33.33	.	.
04	9	27	7.41	40.74	37.04	11.11	3.70
03	18	54	.	18.52	59.26	16.67	5.56
02	8	24	.	12.50	75.00	12.50	.
01	1	3	.	.	100.00	.	.

Table 2c

Site L Systematic Monitoring
Total Reader Population

Performance by Score Point

			Reader Agreement with GED Essay Score				
			- 2 Points	- 1 Point	Same Score	+ 1 Point	+ 2 Points
			Percent	Percent	Percent	Percent	Percent
Score	Essays	Readings					
06	1	3	.	.	100.00	.	.
05	3	9	11.11	33.33	33.33	22.22	.
04	9	27	3.70	11.11	40.74	40.74	3.70
03	18	54	.	18.52	31.48	38.89	11.11
02	8	24	.	4.17	50.00	41.67	4.17
01	1	3	.	.	66.67	33.33	.

References

Breland, H. M. (1986, April). Study objectives and outcomes of reliability and predictive validity analyses. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Ebel, R. L. Estimation of the reliability of ratings. In W. A. Mehrens & R. L. Ebel (Eds.), Principles of educational and psychological measurement. Chicago, IL: Rand McNally & Co. 1967, chap. 17.

General Educational Development Testing Service. (1984). GED tests specifications committee report. Washington, DC: American Council on Education.

General Educational Development Testing Service. (1985a). The relationship between scores on the GED writing skills test and on direct measures of writing (Research Studies No. 6). Washington, DC: American Council on Education.

General Educational Development Testing Service. (1985b). Adding an essay to the GED writing skills test: reliability and validity issues (Research Studies No. 7). Washington, DC: American Council on Education.

General Educational Development Testing Service. (1985c). The 1988 tests of general educational development: A preview. Washington, DC: American Council on Education.

General Educational Development Testing Service. (1986, January). GED essay topics: Some specifications . . . and an invitation. GED Items. p. 3.

Nunnally, J. C. (1978). Psychometric Theory (2nd ed., p. 250). New York: McGraw-Hill.

Patience, W. P. & Swartz, R. Essay score reliability: Issues in and methods of reporting GED Writing Skills scores. Paper presented at the 1987 annual meeting of the National Council on Measurement in Education, Washington, DC, 1987.

Spille, H. A. (1984). Summary of surveys concerning future GED tests (phase 2). (General Educational Development Testing Service Memorandum No. 31). Washington, DC: American Council on Education.

HOW THE GED ESSAY IS SCORED

GED Testing Service

Holistic Scoring

Each paper written for the GED Writing Skills Test is scored by two trained readers. The readers score the papers "holistically;" that is, the piece of writing is evaluated on the basis of its *overall* effectiveness. The salient features of holistic scoring are well described by Conlan (1976):

> The basic assumptions of the holistic reading are that each of the factors involved in writing skill is related to all the others and that no one factor can be separated from the others. Readers must judge each essay as a whole; they must read each paper for the impression its totality makes. A misspelled word, a comma splice, a sentence fragment, a misplaced modifier should carry no great weight in scoring a paper. The candidate is entitled to make some mistakes: he or she is writing hurriedly in a tense situation, without recourse to a dictionary, without the customary time for deliberation. If readers read each paper as a whole, then, they are better able to judge the competence of the writer. If the paper is poorly written, that will be part of the reader's first impression; there is no need to analyze it word by word to decide that it is badly done. If the paper contains mistakes because the writer, though capable, has been forced to hurry, the reader will judge the general quality of the paper most effectively by reading it as a whole. The motto of the reading might well be "Read quickly and judge. Do not reread."

The Six Point Scale

The scale used by each reader in the holistic evaluations ranges from one (low) to six (high). The two readers' scores are then added, resulting in a range of scores from two to twelve. If the two readers' scores differ by more than one point, the paper is scored a third time. The total score for papers scored by three readers is twice the average of the three scores.

Because the six point scale is an even-numbered scale, there is no midpoint. The lack of a midpoint forces readers away from a natural tendency to drift towards the middle. With an even-numbered scale, readers must decide whether each paper belongs in the upper half or the lower half of the papers being scored. Readers can make this decision only if they have some awareness of the total range of writing ability represented by the papers; this is provided by sample papers in conjunction with a descriptive scoring guide.

How the Standards are Defined

The standards for an essay scoring session are defined by the GED essay scoring guide and by sample papers illustrating the different points on the scoring scale. These standards were originally established by examining papers written by a national sample of high school seniors and will be maintained by following the same process for each topic that is used in the GED Writing Skills Test. A group of highly experienced readers reads papers in the pool of high school essays to establish the range of abilities demonstrated by the norming population of (high school) students. After they have selected papers which represent the full range of ability, the readers assign the top score to the best of the papers, the lowest score to the weakest papers, and corresponding scores to other papers throughout the range. The readers' judgments during this process are made independently, and the papers that ultimately serve as samples include only those on which the the readers' scores agree. These sample papers represent the full range of abilities demonstrated by a national sample of high school seniors and thus serve as standard setters, or "rangefinders."

The GED Essay Scoring Guide

The GED essay scoring guide is a further articulation of these standards and accompanies the rangefinders in training sessions. Because it was developed inductively, the scoring guide is descriptive rather than prescriptive. That is, after the readers selected the rangefinders, they attempted to describe, in general terms, the characteristics of papers at different points on the scoring scale. While the sample papers illustrate standards only for the specific topic on which they are written, the scoring guide defines the characteristics papers should exhibit regardless of the topic the paper is written on. Although there is ample research demonstrating that different topics exhibit different strengths and weaknesses of a writer, a generic guide can be used if topics are developed with the scoring guide in mind. All topics developed for use in the GED Essay test are of the same rhetorical type and are of approximately equal length, reading level, and format. In addition, all potential topics are field tested and scored by trained readers, using the GED scoring guide to define standards. If readers find that papers written on a potential topic cannot be scored using the scoring guide, or that papers consistently

exhibit traits other than those described in the guide, the topic is rejected. A further check on the match of topic to scoring guide is made through a statistical analysis wherein the mean and variance of examinee scores on potential topics must approximate the mean and variance of the anchor topic for which the scoring guide was written.

Training of Readers

While the specific procedures for training readers and conducting the essay scoring session depend upon the number of papers to be scored and the number of readers participating, many of the basic principles remain the same regardless of the size of the effort. Readers are usually trained prior to or at the beginning of the essay scoring session. After an introduction to the principles of holistic scoring provided by a chief reader, readers are given the topic, the scoring guide, and the rangefinders, which include at least one paper at each point on the score scale. After a discussion of the tasks required by the topic, and a review of the qualities enumerated on the scoring guide, readers are asked to read the entire set of rangefinders, evaluating them swiftly on the basis of an overall impression and ranking them from best to poorest. Because the set of rangefinders includes more papers than there are points on the score scale (i.e., 8 papers for a six point scale), readers are instructed that they must use the entire range of scores. This is an important instruction in the training because many readers are reluctant to award scores at the top or at the bottom of the score scale. If readers are directed to give the top score to the best paper in the set of rangefinders, they are more likely to acknowledge that a "6" paper does not—and need not—represent perfection.

In a large group training session, the chief reader calls for a show of hands to indicate how readers scored the rangefinders. Ideally, all of the readers will have scored each of the rangefinders identically. In some readings, involving highly experienced readers, this ideal is realized. However, with inexperienced readers there are often substantial differences among readers' scores at this point. To help resolve those differences, readers designated as "table leaders" conduct discussions among the four to six readers at their tables to attempt to bring each individual to a point of consensus with the group. At the heart of the holistic scoring process is the necessity for each reader to be willing to accept—and in some cases consciously shift to—the standards defined in the scoring guide and adopted by the group. Following the discussion of the rangefinders, additional work is done towards this goal through the scoring of more sample papers, distributed singly or in sets of two or three. The scoring and discussion of sample papers continues until the entire group of readers begins to show a consensus in their scoring. At this point, the training period ends and the actual scoring of essays begins, although table leaders may continue to work with individual readers if necessary.

Reliability and Stability

The goals of an essay scoring session are inter-rater reliability and reading "stability." Inter-rater reliability is the degree to which readers agree with each other; generally, the fewer papers which require a third reading, the greater the inter-rater reliability, though this index is affected by other variables as well. Reading stability is the degree to which papers are scored according to the fixed standards described in the scoring guide. This is a particularly important feature of an essay scoring session because the standards for scoring GED essays must remain fixed, regardless of when the essay is administered, where it is scored, or what specific procedures were used in the scoring session itself. It is important to note that a high degree of inter-rater reliability does not ensure reading stability. In short, just because readers are agreeing with each other on essay scores does not mean that they are assigning the scores according to the standards defined on the scoring guide.

To achieve these two goals, the reinforcement of scoring standards continues well after the initial training session is over. As the readers score papers, a table leader (in a small reading) selects scored papers at random to verify that the scoring is consistent with the definitions in the scoring guide. In cases of a disagreement, the chief reader or table leader discusses the paper with the reader. The monitoring process continues throughout the entire scoring session, and, in large readings, the chief reader periodically reviews the scoring of the table leaders.

When scoring large numbers of papers, readers generally read for periods from 30 to 45 minutes at a time and then take breaks. At the beginning of each reading period, additional sample papers are distributed and scored as an additional check on reader accuracy. Generally, readers score between 30 and 50 papers per hour, depending on various factors such as the reader's experience, the length of the papers, and the time of day. Readers are encouraged to read swiftly because a slow, deliberate reading of a paper tends towards analytical, rather than holistic, evaluation of writing. But readers are frequently reminded that accuracy is far more important than speed. Through this system of checks and re-checks, some assurance is gained that all readers are scoring according to the standards defined by the scoring guide and not according to one defined by each individual.

Scoring Method

Considerable attention was given to selecting the best possible method for scoring the writing samples obtained from the high school students. Holistic scoring was selected as the optimal method for this study for several reasons. First, holistic scoring is a quick and relatively simple procedure, so it is an efficient method for scoring large numbers of papers. Because the GED Testing Service wished to consider the feasibility of essay scoring as one of the research questions, holistic scoring was a technique preferable to either analytic or primary trait scoring, both of which demand of essay readers more training and time per essay. Second, to answer the primary research question involving the correlation between essay and multiple choice scores, high degrees of inter-reader reliability were required, and holistic scoring offered the most accessible method of attaining that reliability (Cooper, 1977). An additional factor in the scoring method decision was the acknowledgement that, if a writing exercise were to be added to the GED Writing Skills test, practical constraints would limit the exercise to a single topic. This limitation further suggested the use of holistic scoring and the combination of a holistic essay score with multiple choice scores to yield a composite writing score. The relative unreliability of a holistic score on a single topic (Charney, 1984; Cooper, 1977) is mitigated appreciably when the score is combined with a more reliable indirect measure of writing (Godshalk, et al., 1966).

Holistic scoring is a method of evaluating a piece of writing based upon a paper's overall effectiveness. Often referred to as "general impression scoring," the method relies upon the idea that the effectiveness of a piece of writing does not necessarily equal the sum of its parts, and that informed essay readers can come to agreement about the worth of a whole piece of discourse while they may disagree about the relative value of its component parts (Conlan, 1976).

Because holistic scoring is essentially a process of rank ordering papers (Cooper, 1977), it is frequently observed that one of the limitations of the scoring process is that the ranking indicated by the scores varies according to the population which comprises the sample. That is, a high school student's paper scored as a "six" in one reading may become a "four" if placed among papers written by college graduates because the paper's rank changed in relation to the universe of papers of which it was a member. To control this drifting of standards, a scoring guide was developed for each topic. Three separate guides were written to reflect the clearly different purposes suggested by each of the topics. Because the guides were created inductively—by describing the common characteristics of the agreed-upon model papers—they could be regarded as descriptive rather than prescriptive.

GED ESSAY SCORING GUIDE

Copyright (c) 1987, GED Testing Service

Papers will show <u>some or all</u> of the following characteristics.

Upper-half papers make clear a definite purpose, pursued with varying degrees of effectiveness. They also have a structure that shows evidence of some deliberate planning. The writer's control of the conventions of Standard Written English (spelling, punctuation, grammar, word choice, and sentence structure) ranges from fairly reliable at 4 to confident and accomplished at 6.

6 The <u>6 paper</u> offers sophisticated ideas within an organizational framework that is clear and appropriate for the topic. The supporting statements are particularly effective because of their substance, specificity, or illustrative quality. The writing is vivid and precise, although it may contain an occasional flaw in the conventions of Standard English.

5 The <u>5 paper</u> is clearly organized with effective support for each of the writer's major points. While the writing offers substantive ideas, it lacks the fluency found in the 6 paper. Although there are some errors, the conventions of Standard English are consistently under control.

4 The <u>4 paper</u> shows evidence of the writer's organizational plan. Support, though adequate, tends to be less extensive or effective than that found in the 5 paper. The writer generally observes the conventions of Standard English. The errors that are present are not severe enough to interfere significantly with the writer's main purpose.

Lower-half papers either fail to convey a purpose sufficiently or lack one entirely. Consequently, their structure ranges from rudimentary at 3, to random at 2, to absent at 1. Control of the conventions of Standard Written English tends to follow this same gradient.

3 The <u>3 paper</u> usually shows some evidence of planning, although the development may be insufficient. The supporting statements may be limited to a listing or a repetition of ideas. The 3 paper often demonstrates repeated weaknesses in the conventions of Standard English.

2 The <u>2 paper</u> is characterized by a marked lack of organization or inadequate support for ideas. The development may be superficial or unfocused. Errors in the conventions of Standard English may seriously interfere with the overall effectiveness of this paper.

1 The <u>1 paper</u> lacks purpose or development. The dominant feature is the absence of control of structure or the conventions of Standard English. The deficiencies are so severe that the writer's ideas are difficult or impossible to understand.

* An asterisk code is reserved for papers that are blank, illegible, or written on a topic other than the one assigned. Because these papers cannot be scored, a Writing Skills Test composite score cannot be reported.

5/87

SAMPLE GED ESSAY TEST

Tests of General Educational Development
Writing Skills, Part Two

Instructions

This is a test to find out how well you write. The test has one question which asks you to present an opinion on an issue or to explain something. In preparing your answer for this question, you should take the following steps:

1. Read all of the information accompanying the question.
2. Plan your answer carefully before you write.
3. Use the blank pages of this test booklet (or scratch paper) to make any notes.
4. Write your answer on the separate answer sheet.
5. Read carefully what you have written and make any changes that will improve your writing.
6. Check your paragraphing, sentence structure, spelling, punctuation, capitalization, and usage, and make any necessary corrections.

You will have 45 minutes to write on the question you are assigned. Write legibly and use a ballpoint pen so that the evaluators will be able to read your writing. The notes you make on the blank pages (scratch paper) will not be scored.

Your composition will be scored by at least two trained evaluators who will judge the paper according to its overall effectiveness. They will be concerned with how clearly you make the main point of your composition, how thoroughly you support your ideas, and how clear and correct your writing is throughout the composition. You will receive no credit for writing on a question other than the one assigned.

SAMPLE TOPIC

The automobile has certainly been responsible for many changes in the United States. Some of these changes have improved our lives and some have made life more difficult or unpleasant.

Write a composition of about 200 words describing the effect of the automobile on modern life. You may describe the positive effects, the negative effects, or both. Be specific, and use examples to support your view.

Refer to the steps on the directions page for more specific instructions about writing your composition.

All topics used with GED examinees during the studies remain secure test questions (for possible use in 1988) which cannot yet be published, nor can the examinees responses. However, we are able to provide the example topic above, which was used with a national sample of high school seniors. The responses printed below were written by high school seniors; the scoring scale used to score the papers is the same as that used to score GED examinees' papers.

GED Essay Topics: Some Specifications...and an Invitation

Essay topics, like all other items to be used on the GED Tests, undergo a rigorous scrutiny before they are included on a final form. The process begins, logically enough, with the writing of "raw" topics that address the test specifications. In the case of the essay topics to be used on the GED Writing Skills test, the following general specifications must be met:

- **The topic must be based upon information or a situation that is general enough to be familiar to most examinees.** For example, a topic calling for a writer's views regarding the effect of automobiles on American life might well be appropriate, while one calling for an opinion of the space shuttle's value to the American economy would fail to meet this specification because it would require specialized knowledge. In short, the topic must offer *access* to as many examinees as possible—ideally, to all.
- **The topic must offer an idea that examinees view as worth writing about.** In this sense, the situation defined by the topic's context should be realistic. It would be a simple matter to construct contexts consisting of hypothetical issues, but these could not be expected to fully *engage* the writers. The purpose of the topic is to elicit a sample of writing which displays the examinee's strengths; the kind of topic most likely to do this is one which presents a situation about which the examinee already has an opinion or some feelings, or can form them readily.
- **The topic should not elicit an overly-emotional response from the writer.** While the situation provided as a prompt should trigger the writer's interest in some way, it should not go too far. The larger the population writing on a topic, the greater the chance that the topic will elicit an emotional response from *someone*. Emotional writing is rarely controlled writing and thus does not present an example of an examinee's best writing skill. For this reason, emotionally-charged issues, such as abortion, are not used as writing prompts.
- **The topic should be clearly stated and contain only the amount of information necessary to provide the prompt for writing.** When examinees finish reading the topic, they should know exactly what they are being asked to write about. The objective is to state the topic clearly enough that students immediately know *what* they are going to write and spend their time working out *how* they are going to write it.

Only topics which meet these specifications become eligible for field-testing with GED examinees and with high school seniors. After topics are field-tested, the papers written on topics are holistically scored by experienced readers who use the GED Essay Scoring Guide (see *GED Items*, II.4). This type of "scoring session" differs significantly from one concerned solely with producing scores for essays. In this "topic selection reading," readers are asked not only to score the papers using the Scoring Guide, but also to evaluate how well the topic is working. Readers evaluate topics according to the following specifications:

- **The topic should elicit papers with characteristics comparable to those described in the essay scoring guide.** If, for example, the best papers written on a field-test topic seem significantly less accomplished than those described in the six category in the scoring guide, the topic may be more difficult for writers than desirable and therefore unusable. In the interests of making all of the topics as equal as possible in difficulty, all papers produced by the topics must conform to the scoring guide. If the papers exhibit characteristics that are clearly different from those described in the scoring guide, the topic will be rejected; the scoring guide will *not* be revised to conform to a new topic.
- **The topic should elicit papers which illustrate the full range of student writing ability.** A good topic allows strong writers to display their skill in writing, yet still allows weak writers access to the question. Because the field tests involve significant numbers of examinees and seniors, the total pool of papers produced for a topic should include essays at all points on the six point scale.
- **The topic should elicit papers which clearly address the question provided.** If papers written on a topic consistently show that writers are unclear about the question asked, or if an unusual number of writers are writing on a topic other than the one asked, the topic is ineffective. A topic which fails to meet this criterion was probably not clearly stated to begin with.
- **The topic should elicit a variety of responses.** To say that all papers must address the topic is not to say that all papers must appear the same. A good topic will yield papers whose ideas mirror the diversity of the population tested. Papers with a wide range of ideas will be less tedious for readers to score, and thus readers are likely to score more accurately. While scoring large numbers of papers inevitably becomes tiresome, a good topic produces papers that engage readers' interest to the extent possible.
- **The topic should elicit fully-developed responses.** A topic which yields a large proportion of incomplete papers may be too demanding for the time allowed. If papers consistently exhibit a shallowness of thought and inadequate development, examinees might not have enough information immediately at hand to write well about the topic.
- **The topic should not elicit an emotional or biased response from readers.** In the same way that care is taken not to trigger emotional reactions from writers, the resulting papers must not fuel preconceptions or biases among readers. Essay readers are urged to evaluate the writing, not the writer or the writer's values, and the scoring process contains numerous checks of this standard. However, readers cannot be expected to remain immune from emotional reactions so an effective topic produces few papers that are likely to set off an emotional response in readers.
- **The topic elicits papers for which readers can readily agree on scores.** Readers should be able to distinguish upper half (6,5,4) from lower half (3,2,1) papers with relative ease. Where these distinctions are blurred, the topic itself may well be at fault. Among the many statistical checks on a topic's performance is the rate at which readers agree on scores for particular papers; a significant rate of disagreement among readers often indicates that the topic is yielding papers that cannot reliably be scored using the standards provided.

(continued)

Topics

At this point, you may be wondering how we ever hope to produce even a single operational topic that passes all of these stipulations (along with a few others relating to topic equating). In the spring of 1985, GEDTS field tested fifteen topics with a national sample of high school seniors, and six of the topics were judged acceptable for use in operational forms. To produce the fifteen for field testing, however, we began with well over a hundred topic ideas.

Revised
April, 1987

ESSAY READER QUALIFICATIONS:

To apply for certification as a GED essay reader, a candidate must possess the following:

* a baccalaureate degree;
* at least two years' total experience teaching English-language arts at the secondary, adult, or post-secondary level;
* the ability to write effectively;
* willingness to accept established essay scoring standards;
* openness to the concepts and principles of holistic scoring; and,
* a demonstrated ability to work well in group situations.

ESSAY READER CERTIFICATION:

To be certified as a GED essay reader, a qualified candidate:

* attend a GEDTS-designed holistic scoring training session; and,
* achieve an acceptable score on a set of reader certification papers provided by GEDTS.

Candidates who qualify for certification will be issued a GED Essay Reader Certificate by the state or province administrator. (Persons currently serving as GED teachers may not participate in the reading of GED examinee papers during an actual scoring session.)

CHIEF READER QUALIFICATIONS:

To apply for certification as a GED Chief Reader, a candidate must:

* meet all essay reader qualifications;
* have demonstrated leadership ability;
* have strong communication skills; and,
* have knowledge of holistic scoring procedures (participation in or leadership of scoring sessions preferred).

CHIEF READER CERTIFICATION:

To be certified as a GED Chief Reader, a candidate must be:

* approved by the state or province administrator;
* trained in holistic scoring procedures in accordance with GEDTS Chief Reader guidelines by attending a GEDTS Chief Reader training session;
* willing to supervise GED holistic scoring sessions; and,
* certified as a GED essay reader.

Candidates who qualify for certification will be issued a GED Chief Reader Certificate by the GED Testing Service.

GENERAL EDUCATIONAL TESTING SERVICE
WRITING SKILLS TEST, PART II

Reader Objectivity Issues

Readers should be aware of and attempt to overcome the potential influence of personal bias on their ratings so that they can score each essay as objectively as possible. Bias can affect one's entire scoring pattern or may occur only in specific instances. An overall pattern such as the tendency to score consistently lower than other readers is generally easier to detect and correct than are idiosyncratic biases. A number of factors can bias or unknowingly influence a reader's perception of a paper in a positive or negative direction.

1. Some are characteristics of the paper itself such as:

   a. handwriting;
   b. lightness or darkness of writing;
   c. neatness;
   d. format (skipping lines, unusual margins, etc.);
   e. presence or absence of a title; and
   f. length of the response.

2. Other factors may be more difficult for a reader to identify since they are personal reactions. Perceptions of a paper's overall quality can be influenced by the readers:

   a. expectations for writing performance that are developmentally or instructionally inappropriate for population being scored (e.g.,

   b. personal reactions to the persona or tone of the composition (e.g., "What an egotist!", "Sounds like a cute kid!"; "Another Erma Bombeck!"; etc.);

   c. reactions to the content of the composition (e.g., "It doesn't say anything new."; "It's boring!"; "This paper is vulgar!"; "Not another paper about an elephant!"; etc.);

   d. idiosyncratic preferences or prejudices in style or usage (e.g., the use of profanity; the use of "etc." or "that is");

   e. deeply ingrained assessment standards from previous holistic scoring experience (e.g., this would be a 4 for ETS, our scoring guide is better, can I use it?);

   f. assumptions about the writer or the testing conditions (e.g., "This examinee obviously has a lot of potential but ran out of time."; etc.). Score only what is on the paper.

GENERAL EDUCATIONAL DEVELOPMENT TESTING SERVICE
WRITING SKILLS TEST, PART II

Guidelines For Holistic Scoring

1. Read and Score. Read for an overall impression of the paper. Record your score immediately. Avoid the temptation to reread the paper to justify your score in terms of specific errors.

2. Red pens down. Holistic readers do not make marks on papers they score. There are two major reasons for this. First, there is not time to mark papers in this fashion. Each paper should be scored within a minute or two. Secondly, marking errors directs too much attention to one trait, while allowing the reader to ignore other positive features.

3. Do not change your score. Avoid second-guessing yourself. Make changes only if you made a recording error.

4. When in doubt, refer to the scoring guide criteria. Do not compare a paper to another paper or order papers in terms of quality. Instead, refer to the scale and your internalized concept of the score points.

5. Be aware that there can be a range of papers within a score point. Just as you have students making A's with 100 points and students making A's with 93 points, there are "high", "medium", and "low" papers within each score point.

6. Avoid the tendency to adjust the number of 6's, 5's, 3's, 2's, and 1's in a given set of papers. You may have an entire packet of 2's or an entire packet of 1's.

7. Be aware of an avoid potential biasing factors. (See Reader Objectivity Issues)

8. Score only what is on the paper. Avoid making assumptions about the writer's potential, why the writer omitted an idea, and so forth.

9. This is only a 45 minute test. Remember, examinees had only 45 minutes to read the prompt, to write a first draft, and to write the final draft. Do not compare the finished product with your writing assignments that students have at least a week to prepare.

Appendix I
The General Educational Development Testing Service
of the AMERICAN COUNCIL ON EDUCATION

Report on Systematic Monitoring of Essay Reading Sites

SITE: L/s

SCALE STABILITY (measured on 2 to 12 point scale)

Certification Criteria
Percent agreement with GEDTS..... 87.5%
(percent of essays scored within 2 points of GEDTS score)

Percent of total essay scores
the same as GEDTS..... 15.0%
within 1 point of GEDTS score..... 53.8%
within 2 points of GEDTS score..... 18.8%

differing by more than 2 points from the GEDTS score. 12.5%

Intraclass Correlation between GEDTS and Scoring Site..... .72

READING RELIABILITY	Site	GEDTS
Number of Essays Scored	40	40
Mean Total Score	7.15	6.40
Standard Deviation	2.16	2.02
Rater Agreement (1)		
Percent of essays with reader scores that are equal	35.0%	100%
differ by 1 point (contiguous)	57.5%	
differ by 2 or more points (discrepant)	7.5%	
Intraclass Correlation of readers' scores	.70	1.0
Reading Reliability (2)	.82	

---

(1) Reader agreement is based on the six-point holistic scale used by each reader.

(2) Since the essay score is on the 2 to 12 point scale, intraclass correlations of the sites' reader scores have been stepped-up by applying the Spearman-Brown formula in order to represent the expected "Reading Reliability".

Appendix I
The General Educational Development Testing Service
of the AMERICAN COUNCIL ON EDUCATION

Report on Monitoring Essay Reading Sites

SITE: L/s

READER REPORT

	Number of Essays	(%) Discrepant Scores		Holistic Score Points					
Total		w/GED	Readers	1	2	3	4	5	6
Site	80	12.5	7.5	1.9	16.9	28.1	30.6	17.5	5.0
GEDTS	80	0.0	0.0	2.5	20.0	45.0	22.5	7.5	2.5
Readers									
001	40	10.0	0.0	5.0	20.0	17.5	30.0	20.0	7.5
GEDTS	40	0.0	0.0	2.5	20.0	45.0	22.5	7.5	2.5
005	40	7.5	7.5	0.0	7.5	35.0	32.5	20.0	5.0
GEDTS	40	0.0	0.0	2.5	20.0	45.0	22.5	7.5	2.5
009	40	7.5	7.5	2.5	32.5	25.0	27.5	10.0	2.5
GEDTS	40	0.0	0.0	2.5	20.0	45.0	22.5	7.5	2.5